NASA TT F-11,391

COMMAND BY SPEECH IN AEROSPACE AUTOMATION

F.Winckel

Translation of "Befehlssprache in Luft- und Raumfahrtautomatik". Wissenschaftliche Gesellschaft für Luft- und Raumfahrt und Deutsche Gesellschaft für Raketentechnik und Raumfahrtforschung, Jahrestagung, Berlin, West Germany, September 14-18, 1964.
11 pp.

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
WASHINGTON, D.C. 20546 DECEMBER 1967

NASA TT F-11,391

COMMAND BY SPEECH IN AEROSPACE AUTOMATION /1*

F.Winckel

ABSTRACT. A discussion of automatic control based on the recognition of phonetic patterns by control equipment is given. The effects of delayed speech feedback are reviewed, and the acoustic properties of speech are described. A formant-coding speech-compression system is examined in which additional linguistic control is achieved by using the statistical dependence of subsequent tones in the sense of a Markov chain. It is shown that the state of the art enables the realization of a command language consisting of about 10 words - the numbers from 0 to 9, for example.

1. Manual Control and Control by Speech

With the increase in complexity of control functions in machines and systems, the requirements to bring these functions into correlation with various laws, in order to process them in part or in totality in electric circuits or on computers, are constantly increasing. This problem belongs to the field of automatic or semi-automatic execution and is treated in the technique of process control. The model of information processing in the human nervous system represents an analog which, as a prototype of nature, furnishes various indications for solving regulation and control-technical problems. Below, we will study the possibility of reducing the manually given control commands in connection with monitoring of the control devices or even with replacing by speech command.

The close correlation between thinking and speaking, which can be considered as a cybernetic speech-thinking circuit, suggests possibilities of direct conversion of thinking into speech or, in the case in question, into speech commands which, as acoustic signals, are able to release control functions converted into patterns. There is no doubt that this method will result in shorter reaction times than possible by manual control, such as operation of pushbuttons. It is true that the delay time for performing a given manipulation is surprisingly short; nevertheless, usually a sequence of such manipulations is involved. Depending on the number of pushbuttons to be operated in connection with different instruments, the prolongation of the reaction time constants will be logarithmically obtained as a selective quantity in bits.

These considerations are basic for the manual control of a manned space capsule or for the cockpits of modern multipurpose aircraft. The visual display of the operating factors and the navigation quantities are not yet coordinated according to synoptic viewpoints, so that the pilot still has to solve a mental coordination problem before executing any act, not even considering the selection

[*] Numbers in the margin indicate pagination in the foreign text.

of the instantaneously required values from the totality of all readings. Unrequired readings are felt as a disturbing factor, thus increasing the noise /2 level for perception and diminishing the required information value for the percipient. The situation is complicated further by the difficulty of assimilating the visual quantities, perceived outside the vehicle, with the displays seen inside the capsule; in addition, this requires accommodation of the eye from distant to close-in vision. For example, at a flying speed of 1100 km/hr, a delay of 0.9 km occurs when shifting the eye from outside to the instruments and back again. In each case, the reaction times given as normal for human subjects will increase manyfold under the existing psychic stresses of multiple aircraft or space-vehicle control problems. Consequently, any manual multiple switching results in a much longer delay time than the operation of a pushbutton when fully rested.

A command by speech would shorten the total delay time, provided that the neuromuscular speech motor function of the subject functions reliably. In the subjective monitoring of the speaking by the individual's own ear - cochleorecurrent loop or control circuit - a delay time of 70 msec is observed for the case of simple sounds (Ref.1), meaning that the signal reaction in the form of sounds to a visual event must be estimated as more than 100 msec. For words instead of simple sounds, the delay time will be 250 msec (Ref.2). However, as mentioned above, any decision selection from numerous observed events has an additional delaying effect. Conversely, the problem complex presents great difficulty in constructing voice-controlled machines as a communication between man and machine; this difficulty lies in the irrationality of the speech phenomenon, as it is true also for various handwritings. Both phenomena form one of the main subjects of present-day research, but no solutions for the strictness of recognition of acoustic and visual structures are in sight. This is due mainly to the individuality of human nature which opposes any reproducible behavior, so that the recognition of conventional speech cannot be subjected to norming solutions. Consequently, if no linear coordinations of linguistic symbols with objectively determinable quantities exist, it would be necessary to use speech systems of logistic types; however, this would require the intercalation of some system, /3 even if only on a mental plane, which again would prolong the reaction time constant. This is further complicated by the difficulty of speech processing since the average person, because of the redundancy of language, is rather negligent in speaking and writing; frequently, symbols respectively phonemes or even whole sequences of these are omitted without interfering with the understanding by the listener. This places further limits on automatic speech recognition, which means that there is not much chance to ever go beyond a phonetic or phonemic transcription of the speech signal.

2. Characteristics of Speech

At first, rather rough and tentative solutions for automatic recognition of writing and speech must be tolerated. Below, we will restrict ourselves to the speech problem in application possibilities to space travel.

When conceiving speech acoustically as a sequence of varying quasi-periodic processes and noises - corresponding to vowels and consonants - it is logical to perform an analytic resolution into physical elementary structures and, in accordance with the invariants contained therein, to form the code of new characters and, for example, to thus release the type bars of a typewriter. This method was used by J. Dreyfus-Graf (Ref.3) on a restricted scale, with some success. An analysis of the approximately periodic processes can be carried out by a Fourier analysis in acoustic spectra, whose changes as a function of time can be recorded (for example) on a visible speech display (Ref.4). An Atlas of all speech sounds gives a very rough approximation of the relations existing between linguistic sounds or phonemes and the frequency analysis; such an Atlas also indicates the large number of variants in sound transitions and in dependence of individual speakers. Specifically, experiments by C.Harris and K.Küpfmüller (Ref.5, 6) indicated that, even at distinct articulation, it is impossible to resolve the speech recorded on magnetic tape into individual sounds by cutting the tape with scissors and then to paste up new words. Such a speech would remain completely unintelligible. This led to the realization that the transitions between two sounds are the main determining factor and the vowel gradation is influenced by the preceding and subsequent consonant. Our slide No.1 shows the 4 width of the toning frequency range of the vowels - the formants F_1 and F_2 - within which characteristic vowels can be recognized, or vice versa the poor accuracy of determinability of the spectrum of the vowel formants.

In addition, the sound identification depends on various parameters such as pitch, sound intensity, and sound duration which represent special significance carriers in the context of speech and may even have an emotional character. It is exactly the latter which must be avoided in any command by speech. Consequently, the sound quality cannot be quantized per se.

Thus, a second and presumably more essential characteristic of speech is required, namely the linguistic correlation which gives the listener additional information on the transition probabilities in the speech sound sequence. At higher stages, the main question is that of forming correlations within the speech structure and over differing time intervals. In any case for speech recognition several physical characteristics cooperate to form one act of recognition, for which the pertaining sound sequence is of basic influence, as demonstrated by experiments of Haskins Laboratories (Ref.7). However, these correlations are not that simple as to permit deriving the linguistic function from the acoustic structure.

Going beyond this point, it can be stated that many information-carrying symbol phenomena are undetectable within the acoustic complex wave and must be taken from the context of further information hints. In addition, the speech perception depends largely on the learning process which starts in the first year of life and leads to a prejudice with respect to classification of the recognized sounds. This results in the peculiarity that the acoustically perceived sounds have no physical relation with the spoken sounds.

The learning process leads to the ability of grouping sounds of different physical characteristics into one class and, on the other hand, of differentiating between physically similar sounds if a certain linguistic category requires this.

3. Analysis-Synthesis Technique /5

In the development of speech recognition devices, considerable preliminary work has been done in communications technique by establishing coding methods for speech with reduced channel capacity during transmission. In this analysis-synthesis telephone principle, the spectrum of the continuously varying speech structure is approximately resolved into its harmonic components or into percentual intensities in successive frequency ranges, after which the components are rectified and smoothed by a 25-cycle low-pass filter. In that case, the intensity fluctuates only with the relatively slow muscular movement of the articulation organ. The slide No.2 shows the wiring diagram of this principle expanded to the Vocoder, as developed by the Bell Telephone Laboratories (Ref.8) and demonstrated for the first time in 1936 by Dudley (Ref.8). The original idea, however, was developed by K.O.Schmidt according to a patent application of 1932. Here, the spectrum is subdivided into channels of 300-cycle bandwidth in the telephone transmission range of about 300 - 3500 cps. The outputs of the analyzer end are transferred to the receiver end and there build up the sound synthesis which, in the loudspeaker, reproduces the reconstructed speech. A corresponding filter set of the synthesis end is controlled by the filter outputs of the analysis end, in which case a buzzer is used as power source for the vowels and a noise generator for the aperiodic consonants. Differentiation between vowels and consonants is carried out by a frequency meter on the pickup side which responds only to periodic processes. This causes a relay on the receiver end to switch from the periodic source to the noise source.

Suitable filter sets, during the further development, were used for visualizing the course of speech structures in the time spectrum (slide No.3) which, in the meantime, has become an important analytical tool known as "visible speech" (Ref.4).

The original Vocoder consisted of ten filters. Practical experience has shown that eight filters as lower limit of intelligibility are sufficient, while stricter requirements would need 18 filters and more. In view of the large ex- /6 penditure for such channel Vocoders, an attempt was made to restrict the number of channels to the most essential formant regions which must be considered as the most important information carriers. This led to the development of so-called formant Vocoders (Ref.9).

Aside from the spectral analysis and spectral synthesis of speech as a physical or technically imaging method, the following characterizations are used as basis for an automatic recognition of speech, which define the speech character more accurately: speech sound or phoneme unit, syllable unit, and word unit. It is true that, with an increase in units,the intelligibility of speech increases but this also leads to an increase in the required memory capacity. Consequently, before each individual planning, the limits of the character inventory must be defined. Let us assume that, of the given inventory of a language (for example the English language), only 75% of the characters (sounds, syllables, words) are used; in that case, 19 sounds 339 syllables or 732 words would be sufficient (Ref.10). However, even a phoneme* is not a quantic component but requires 12

* "Phonemes" are the variants of lowest order in speech-sound communication,i.e., symbols that do not consist themselves of symbols (according to Meyer-Eppler).

characteristics for its identification (Ref.11) whose number, however, can be slightly reduced in the various languages. Several characteristic circuits will be required for a phoneme recognizer while a word recognizer requires a number of phoneme recognizers equal to the phonemes occurring in the word in question; these recognizers must be connected in series.

The original attempts at an automatic speech recognition used the described filter analysis. J.Dreyfus-Graf in Geneva (Ref.12) was the first to attempt a conversion of speech into written characters, started from a spectral analysis in six filter channels within the 80 - 3800 cycle range; the output of these channels, after rectification and smoothing, was fed to the control of a sixtuple coil system where it controlled a recorder in a complex curve such that the written characters approximately resembled letters. This vector recorder, controllable in six directions, was later replaced by a matrix of differential relays that were tripped in accordance with the speech composition and operated the keys of a typewriter in the respective combination. In addition to this vowel de- /7
termination, the envelope of the wave train was used for recognition of consonants, in which case the analysis was made in a set of four filters in the 2 - 64 cycle range. The resultant components were designated by Dreyfus-Graf as "subformants".

A difficulty - as also in other speech recognizers - consisted in determining the instant at which one sound ends and the other begins since, in reality - as mentioned before - dragging transitions might occur. The speech course can be periodically clocked, but matching of the clock to a given phoneme sector is preferable, which can be obtained from the envelope of the acoustic pressure curve by means of gating circuits.

The Dreyfus-Graf method is restricted to a certain number of phonetic characters. In addition, these characters are repeated once or several times, in accordance with the duration of enunciation, such as

wwoord   beccomes   deeed   (word becomes deed).

Another machine translation of speech to typewriter was developed by H.F.Olson at the RCA Laboratories (Ref.13).

In other methods of speech recognition, the spectral analysis in accordance with the analysis-synthesis technique again forms the starting point. Here, the interpretation of the spectral components is done in a matrix for the required character or pulse formation for control purposes. The number of obtainable phonemes is always limited by the size of the matrix, which also restricts the number of available words formed from these. A matrix of 10 × 10 elements will just about constitute a tolerable circuitry outlay. This is the reason for the fact that many research centers have not gone beyond the reproduction of digits from 1 to 10.

As a typical example, let us mention the unit developed by the Bell Telephone Laboratories which has undergone several developmental stages since 1952 (Ref.14) and finally was demonstrated under the name of "Audrey". As shown in slide No.4, the speech here is spectrally resolved by a set of 300-cycle filters in the range up to 3000 cps, after which the filter outputs are rectified and smoothed. The resultant spectrum is compared, in a matrix of potentiometers, with stored samples of phonemes; by operating one of ten relays, the phoneme to be displayed is selected from a maximum condition. Essentially, this represents a process of cross correlation.

In a subsequent stage, words are formed from the phonemes, in this case /8
the digits 1 - 10. This is again followed by a comparison with the previously recorded and stored words. A criterion for recognition is the duration of the individual phoneme groups in the speech and the order sequence of the phonemes. This comparison is again carried out by cross correlation detection.

This recognizer operates satisfactorily if it is adjusted to a certain person, but is still far from a general applicability. In the laboratory, it has been possible to convert the voltage values for the numerals to the pulse sequences of the telephone dial and thus to obtain a number selection by voice command.

Based on this fundamental principle, Dudley (Bell Telephone Laboratories) developed a phonetic Vocoder (recognition Vocoder) in which the pulses, derived over the filter set and the matrix, are used as a new speech code, for example binary signals similar to the signals on a teletype (Ref.15). These signals can be stored on punch tape or can be used directly at the receiver for operating a typewriter. The required bandwidth for transmission is only 50 cycles, i.e., one sixth of the bandwidth required for the channel Vocoder; incidentally, this is the same bandwidth needed for telegraphic transmission of the same text. Experiments were made with an alphabet of four consonants and six vowels.

At M.I.T. (Ref.16), the Whirlwind I computer is being used for speech recognition. The signal resolution of the speech takes place in a set of 35 bandpass filters with subsequent rectification and smoothing, followed by scanning of the values 180/sec with a rotary switch. These "samples" are digitally converted, for analysis by the computer or for storage on magnetic tape. The computer input is fed at a rate of 96,300 bits/sec. Based on computing the relatively high-frequency components, a differentiation is made between "sonorant and nonsonorant" segments, i.e., whether a formant structure does or does not exist. The determination of the two first formants is shown in slide No.5.

Starting from the elementary symbol characters of the language and using /9
the distinctive characteristics by Jacobson and Halle (Ref.11), one finally arrives at a design of phoneme recognizers that perform a classification by recurrent dichotomy, as shown in slide No.6. The supplied speech signal is segmented by characteristics separators, a process repeated in successive steps by further separators. A total of n successive binary separations lead to an identification of 2^n phonemes. In the design of a recognizer developed by Wiren and Stubbs (Ref.17), the sequence of decision pairs proceeds as follows: sonant - surd, plosives - spirants, nonturbulent (sonant) - turbulent sounds, vowel - vowel-like consonants, and light - dark. For example, the differentiation between sonant and surd is done by means of a low-pass filter with a limiting frequency of 175 cps, carrying the instruction that, on exceeding a certain threshold value, the reading will show "sonant". If such reading does not appear, this is a sign for "surd". By such circuits and others of a similar type, the other characteristics can be binarily differentiated. This device represents a highly interest-

ing testing tool for research on the physical nature of speech sounds.

The speech recognizer by Fry and Dems (Ref.18) represents a higher stage of development in so far as a linguistic interpretation is added to the device, for acoustical analysis. The acoustic spectral analysis is carried out with a set of 18 filters of 1/3 octave spacing over a range of 160 to 8000 cps. As usually, rectification and smoothing of the filter outputs is done with time constants of 10 - 1 msec, descending from low to high frequencies. These outputs which, in their sequence, represent the envelope of the spectrum are compared with each of the stored spectral samples of phonemes. This comparison is made by means of a multiplication, in which the maxima of the multiplication products are selected (slide No.7). Wherever difficulties occur in the individual sounds, a further identification is made, for example using the differing sound duration in the case of plosives and the separation of |s| from |z|.

/10 The additional linguistic control is obtained by interpreting the statistical dependence of successive phonemes in the sense of a Markov chain. In this manner, the transition probability for each previously identified phoneme is determined. Thus, two memories are necessary, namely one that records the transition probabilities of all existing phonemes which for technical reasons naturally are restricted, as "diagram frequencies", and another memory which, operating as a temporary memory, remembers only the just preceding phoneme. These memories consist of a matrix of potentiometers whose outputs are compared with the potentials of the voice-acoustically derived potentials, resulting again in maxima as the values of greatest probability.

The principal wiring diagram is shown in slide No.8. With this device, a correct phoneme recognition of 70% in the best case is reached, while the accuracy for word recognition is only 45%.

This brief and relatively incomplete survey over possible circuitry for automatic speech recognition indicates that we are far from recognition of continuous speech and even of words of a more extensive inventory but that a command speech of about 10 words (for example, digits from 0 to 9) can be realized in practice. This is no doubt sufficient for giving the most necessary and most frequent switch commands in a space capsule or an aircraft cockpit. If, in addition, the pilot is in voice connection with the crew members and the ground station, a switch-over to command speech would be required. However, this is done more rapidly than selecting from a number of 10 pushbuttons. The eye is not deflected from the display units on the control panel and the physical stress during speaking is less than that produced by frequently changing manual pushbutton selection, a fact well known from operating an automatic telephone dial compared with the single-pushbutton system.

It is known that, in the physical isolation of one-man space flights, any occupation is useful for psychic reasons; however, it is of importance to make use of speech since this is the best deterrent for possible introversion of the one-man crew and since a defense mechanism is set up by the necessary speaking position.

REFERENCES

1. Lee, B.S.: Effects of Delayed Speech Feedback. J. Acoust. Soc., Vol.22, pp.824-826, 1950.
2. Küpfmüller, K.: Third ICA Congress for Acoustics, Stuttgart 1959; Amsterdam 1961, Verlag Elsevier.

2a. Winckel, F.: Cybernetic Functions in Vocalization and in Speaking (Kybernetische Funktionen bei der Stimmgebung und beim Sprechen). Phonetica, Vol.9, pp.108-126, 1963.

3. Dreyfus-Graf, J.: Phonetograph and Subformants (Phonétographe et subformants). Techn. Mitt. PTT, Vol.35, pp.41-58, 1957.
4. Potter, R.X., Kopp, G.A., and Green, H.C.: Visible Speech, New York, 1947.
5. Harris, C.M.: Study of Building Blocks in Speech. J. Acoust. Soc., Vol.25, pp.962-969, 1953.
6. Küpfmüller, K.: Speech Synthesis from Sounds. Nachr-tech. Fachber.3, 1956.
7. Delattre, P.C. and Coll.: Acoustic Loci and Transitional Cues for Consonants. J. Acoust. Soc., Vol.27, No.4, p.769, 1955.
8. Dudley, H.: Remaking Speech. J. Acoust. Soc., Vol.11, pp.169-177, 1939.
9. Flanagan, J.L.: Development of a Formant-Coding Speech Compression System. J. Acoust. Soc., 1956.
10. Olson, H.F. and Belar, H.: Phonetic Typewriter. J. Acoust. Soc., Vol.28, pp.1072-1081, 1956.
11. Jacobson, R. and Halle, M.: Fundamentals of Language. s'-Gravenhage, 1956.
12. Dreyfus-Graf, J.: Phonetograph, its Present and Future (Phonétographe: Présent et Futur). Bull. Tech. PTT, p.160, 1961.
13. Olson, H.F. and Belar, H.: Printout System for the Automatic Recording of Spoken Syllables. J. Acoust. Soc., Vol.34, p.166, 1962.
14. Dudley, H. and Balashek, S.: Automatic Recognition of Phonetic Pattern in Speech. J. Acoust. Soc., Vol.30, pp.721-732, 1958.
15. Dudley, H.: Phonetic Pattern Recognition Vocoder. J. Acoust. Soc., Vol.30, pp.733-738, 1958.
16. Hughes, C. and Halle, M.: On the Recognition of Speech by Machine. Information Processing. UNESCO Paris, 1960.
17. Wiren, J. and Stubbs, H.L.: Electronic Binary Selection for Phoneme Classification. J. Acoust. Soc., Vol.28, pp.1082-1091, 1956.
18. Dems, P.: Design and Operation of the Mechanical Speech Recognizer at University College London. J. Brit. Radio Eng., p.219, 1959.

Translated for the National Aeronautics and Space Administration by the O.W.Leibiger Research Laboratories, Inc.